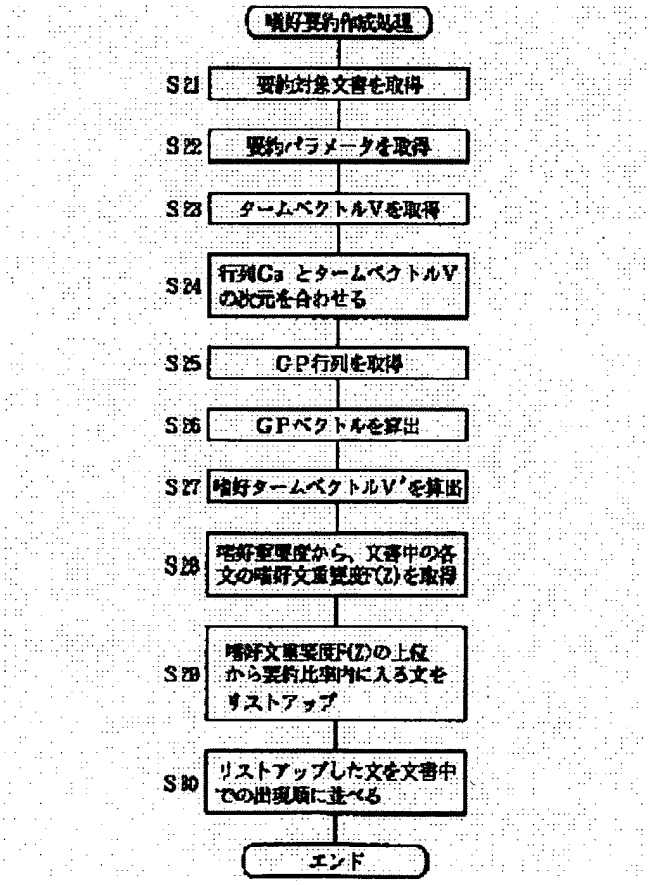
DOCUMENT PROCESSOR, STORAGE MEDIUM STORING DOCUMENT PROCESSING PROGRAM AND DOCUMENT PROCESSING METHOD

Patent number:	JP11045289
Publication date:	1999-02-16
Inventor:	NOMURA NAOYUKI
Applicant:	JUST SYST CORP
Classification:	
- international:	G06F17/30; G06F17/27
- european:	
Application number:	JP19970218230 19970728
Priority number(s):	JP19970218230 19970728

Abstract of **JP11045289**

PROBLEM TO BE SOLVED: To provide a document processor capable of preparing a summary based on the preference of a user such as a utilization purpose or the like, a storage medium storing a document processing program and a document processing method. SOLUTION: A key word and the importance are obtained from the contents of a processing document in the past and a GP(group personalizing) matrix for which one of the plural users and the key word is turned to a row, the other is turned to a column and the importance of the respective key words to the respective users is turned to an element value is obtained. Important words (a), (b),... from a summary preparation object document and the importance from the appearing frequency or the like are obtained, a term vector V for which the importance is an element is shifted by the GP matrix and a preference term vector V' is obtained. Preference important sentences F(Z) are extracted from the summary preparation object document based on the element (=preference importance) of the preference term vector V', arranged in an appearing order in the summary preparation object document and turned to a preference summary.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平11－45289

(43)公開日　平成11年(1999)２月16日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ	17/30	Ｇ０６Ｆ	15/401	３２０Ａ
	17/27		15/20	５５０Ａ
			15/40	３７０Ａ
			15/403	３４０Ａ

審査請求　未請求　請求項の数７　ＦＤ　（全 12 頁）

(21)出願番号　特願平9－218230

(22)出願日　平成９年(1997)７月28日

(71)出願人　390024350
株式会社ジャストシステム
徳島県徳島市沖浜東３－46

(72)発明者　野村　直之
徳島県徳島市沖浜東３丁目46番地　株式会社ジャストシステム内

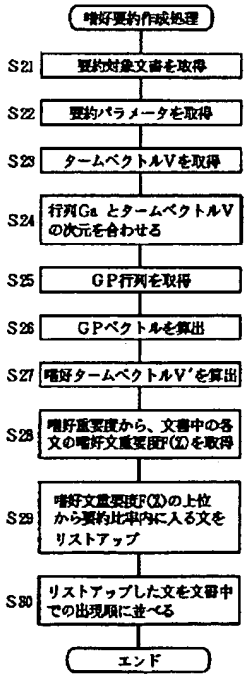
(74)代理人　弁理士　川井　隆　（外１名）

(54)【発明の名称】　文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法

(57)【要約】

【課題】　本発明は、利用目的等のユーザーの嗜好を踏まえた要約の作成が可能な、文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法を提供すること。

【解決手段】　過去の処理文書の内容からキーワードとその重要度を取得し、複数のユーザーとキーワードとの一方を行、他方を列として前記各ユーザーに対する各キーワードの重要度を要素値とするＧＰ行列を取得する。要約作成対象文書から重要語ａ，ｂ，…と、その出現頻度等からの重要度を取得し、この重要度を要素としたタームベクトルＶを、ＧＰ行列によってシフトさせ、嗜好タームベクトルＶ’を取得する。嗜好タームベクトルＶ’の要素（＝嗜好重要度）をもとに要約作成対象文書から嗜好重要文Ｆ（Ｚ）を抽出し、要約作成対象文書における出現順に並べて、嗜好要約とする。



【特許請求の範囲】
【請求項1】　複数の文よりなる文書を取得する文書取得手段と、
前記文書取得手段により取得された前記文書から重要語句とその重要度を取得する重要語句抽出手段と、
前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択する嗜好重要部分選択手段と、
前記嗜好重要部分選択手段により選択された嗜好重要部分に基づいて前記文書の要約を作成する嗜好要約作成手段とを具備することを特徴とする文書処理装置。
【請求項2】　前記重要語句抽出手段は、
前記文書取得手段により取得された前記文書から前記重要語の候補語句とその重要度を取得する候補語句取得手段と、
ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと各ユーザーの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得手段と、を有し、
前記嗜好取得手段により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記候補語句取得手段により取得された候補語句の重要度をシフトさせた重要度から前記重要語句を抽出し、
前記嗜好重要部分選択手段は、前記重要語句とその重要度により前記嗜好重要部分を選択することを特徴とする請求項1に記載の文書処理装置。
【請求項3】　前記重要語句抽出手段は、前記文書取得手段により取得された前記文書から前記重要語の候補語句とその重要度を取得して、前記候補語句の重要度により前記重要語句を抽出し、
前記嗜好重要部分選択手段は、ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと複数のユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得手段を有し、前記嗜好取得手段により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記重要語句抽出手段により取得された重要語句の重要度をシフトさせた重要度により前記重要部分を選択することを特徴とする請求項1に記載の文書処理装置。
【請求項4】　複数の文よりなる文書を取得する文書取得機能と、
前記文書取得機能により取得された前記文書から重要語句とその重要度を取得する重要語句抽出機能と、
前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択する嗜好重要部分選択機能と、
前記嗜好重要部分選択機能により選択された嗜好重要部分に基づいて前記文書の要約を作成する嗜好要約作成機能とをコンピュータに実現させるためのコンピュータ読みとり可能な文書処理プログラムが記憶された記憶媒体。
【請求項5】　前記重要語句抽出機能は、
前記文書取得機能により取得された前記文書から前記重要語の候補語句とその重要度を取得する候補語句取得機能と、
ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと各ユーザーの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得機能と、を有し、
前記嗜好取得機能により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記候補語句取得機能により取得された候補語句の重要度をシフトさせた重要度から前記重要語句を抽出し、
前記嗜好重要部分選択機能は、前記重要語句とその重要度により前記嗜好重要部分を選択することを特徴とする請求項4に記載した文書処理プログラムが記憶された記憶媒体。
【請求項6】　前記重要語句抽出機能は、前記文書取得機能により取得された前記文書から前記重要語の候補語句とその重要度を取得して、前記候補語句の重要度により前記重要語句を抽出し、
前記嗜好重要部分選択機能は、ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと複数のユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得機能を有し、前記嗜好取得機能により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記重要語句抽出機能により取得された重要語句の重要度をシフトさせた重要度により前記重要部分を選択することを特徴とする請求項4に記載した文書処理プログラムが記憶された記憶媒体。
【請求項7】　複数の文よりなる文書を取得し、
取得された前記文書から重要語句とその重要度を取得し、
前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択し、
選択された前記嗜好重要部分に基づいて前記文書の要約を作成することを特徴とする文書処理方法。
【発明の詳細な説明】
【0001】
【発明の属する技術分野】本発明は、文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理

方法に関し、更に詳細には、利用目的等のユーザーの嗜好を踏まえた要約の作成に関する。

【0002】

【従来の技術】従来、書籍、論文、報告書等の各種の文書に対し、要約（抄録を含む）の自動作成処理をコンピュータを用いて行うことが行われている。文書の自動要約については、例えば、「全文情報からの意味的情報の抽出と加工」（情報処理学会第38回全国大会予稿集、第222頁；1989年）で提案されている。この方法では、まず文書中の重要語を字種や動詞等の情報から抽出し、さらに重要語の出現頻度から最重要語を取得する。次に重要語と最重要語が出現するか否かから重要文を取得することで、自動的に要約を作成することが可能になる。また、文章の段落の性質を反映させることで、より正確に要約を作成する特開平3－191475号公報に記載された方法等も提案されている。

【0003】

【発明が解決しようとする課題】しかし、同一の文書でも、例えば営業用や技術資料用等の利用目的その他のユーザーの嗜好が異なると、文書における重要部位等に差異が生じる。そして、上述のような従来の文書処理によって要約を作成しても、ユーザーの嗜好を踏まえた要約を得ることはできない問題点がある。

【0004】本発明は、上述のような課題を解決するためになされたもので、利用目的等のユーザーの嗜好を踏まえた要約自動作成処理文書処理を行うことのできる文書処理装置、文書処理プログラムを記憶した記憶媒体、及び文書処理方法を提供することを目的とする。

【0005】

【課題を解決するための手段】請求項1に記載の発明は、複数の文よりなる文書を取得する文書取得手段と、前記文書取得手段により取得された前記文書から重要語句とその重要度を取得する重要語句抽出手段と、前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択する嗜好重要部分選択手段と、前記嗜好重要部分選択手段により選択された嗜好重要部分に基づいて前記文書の要約を作成する嗜好要約作成手段と、を具備する文書処理装置を提供することにより、上記目的を達成する。請求項2に記載の発明は、請求項1に記載の文書処理装置において、前記重要語句抽出手段は、前記文書取得手段により取得された前記文書から前記重要語の候補語句とその重要度を取得する候補語句取得手段と、ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと各ユーザーの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得手段と、を有し、前記嗜好取得手段により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記候補語句取得手段により取得された候補語句の重要度をシフトさせた重要度から前記重要語句を抽出し、前記嗜好重要部分選択手段は、前記重要語句とその重要度により前記嗜好重要部分を選択することを文書処理装置を提供することにより、上記目的を達成する。請求項3に記載の発明は、請求項1に記載の発明において、前記重要語句抽出手段は、前記文書取得手段により取得された前記文書から前記重要語の候補語句とその重要度を取得して、前記候補語句の重要度により前記重要語句を抽出し、前記嗜好重要部分選択手段は、ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと複数のユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得手段を有し、前記嗜好取得手段により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記重要語句抽出手段により取得された重要語句の重要度をシフトさせた重要度により前記重要部分を選択する文書処理装置を提供することにより前記目的を達成する。請求項4に記載の発明は、複数の文よりなる文書を取得する文書取得機能と、前記文書取得機能により取得された前記文書から重要語句とその重要度を取得する重要語句抽出機能と、前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択する嗜好重要部分選択機能と、前記嗜好重要部分選択機能により選択された嗜好重要部分に基づいて前記文書の要約を作成する嗜好要約作成機能とをコンピュータに実現させるためのコンピュータ読みとり可能な文書処理プログラムが記憶された記憶媒体を提供することにより上記目的を達成する。請求項5に記載の発明は、請求項4に記載の記憶媒体において、前記重要語句抽出機能は、前記文書取得機能により取得された前記文書から前記重要語の候補語句とその重要度を取得する候補語句取得機能と、ユーザーの嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと各ユーザーの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得機能と、を有し、前記嗜好取得機能により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記候補語句取得機能により取得された候補語句の重要度をシフトさせた重要度から前記重要語句を抽出し、前記嗜好重要部分選択機能は、前記重要語句とその重要度により前記嗜好重要部分を選択する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成する。請求項6に記載の発明は、請求項4に記載の記憶媒体において、前記重要語句抽出機能は、前記文書取得機能により取得された前記文書から前記重要語の候補語句とその重要度を取得して、前記候補語句の重要度により前記重要語句を抽出し、前記嗜好重要部分選択機能は、ユーザー

の嗜好を表す複数のキーワードの重要度を要素値とする嗜好ベクトル、または、複数のユーザーと複数のユーザーそれぞれの嗜好を表す複数のキーワードとの一方を行、他方を列として前記各ユーザーに対する前記各キーワードの重要度を要素値とするＧＰ行列、を取得する嗜好取得機能を有し、前記嗜好取得機能により取得された前記嗜好ベクトルまたは前記ＧＰ行列を用いて、前記重要語句抽出機能により取得された重要語句の重要度をシフトさせた重要度により前記重要部分を選択する文書処理プログラムが記憶された記憶媒体を提供することにより前記目的を達成する。請求項７に記載の発明は、複数の文よりなる文書を取得し、取得された前記文書から重要語句とその重要度を取得し、前記重要語句に基づいて前記文書からユーザーの嗜好を反映した嗜好重要部分を選択し、選択された前記嗜好重要部分に基づいて前記文書の要約を作成する文書処理方法を提供することにより前記目的を達成する。

【０００６】

【発明の実施の形態】以下、本発明の文書処理装置、文書処理プログラムが記憶された記憶媒体、及び文書処理方法の好適な実施の形態について、図１から図７を参照して詳細に説明する。

（１）実施形態の概要

本実施形態では、過去の処理文書の内容からキーワードとその重要度を取得し、複数のユーザーとキーワードとの一方を行、他方を列として前記各ユーザーに対する各キーワードの重要度を要素値とするＧＰ行列を取得する。要約作成対象文書から重要語ａ，ｂ，…と、その出現頻度等からの重要度ｇ（ｐ），ｇ（ｑ），…を取得し、重要度を要素としたタームベクトルＶ＝（ｇ（ｐ），ｇ（ｑ），…）を、ＧＰ行列によってシフトさせ、嗜好タームベクトルＶ’を取得する。嗜好タームベクトルＶ’の要素（＝嗜好重要度）ｇ’（ｐ），ｇ’（ｑ），…をもとに要約作成対象文書から嗜好重要文を抽出し、要約作成対象文書における出現順に並べて、嗜好要約とする。

【０００７】（２）実施形態の詳細

図１は、本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムを記憶した記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。この図１に示すように、文書処理装置（コンピュータ）は、装置全体を制御するための制御部１１を備えている。この制御部１１には、データバス等のバスライン２１を介して、入力装置としてのキーボード１２やマウス１３、表示装置１４、印刷装置１５、記憶装置１６、記憶媒体駆動装置１７、通信制御装置１８、および、入出力Ｉ／Ｆ１９、および、文字認識装置２０が接続されている。制御部１１は、ＣＰＵ１１１、ＲＯＭ１１２、ＲＡＭ１１３を備えている。ＲＯＭ１１２は、ＣＰＵ１１１が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。

【０００８】ＲＡＭ１１３は、ＣＰＵ１１１にワーキングメモリとして使用されるランダムアクセスメモリである。このＲＡＭ１１３には、本実施形態による嗜好要約処理を行うためのエリアとして、対象文書格納エリア１１３１、要約パラメータ格納エリア１１３２、重要語格納エリア１１３３、タームベクトル格納エリア１１３４、行列格納エリア１１３５、嗜好タームベクトル格納エリア１１３６、要約格納エリア１１３７、その他の各種エリアが確保されるようになっている。

【０００９】対象文書格納エリア１１３１には、嗜好要約作成の対象となる文書が格納される。要約パラメータ格納エリア１１３２には、操作者からの入力等により取得された要約パラメータの値または後述のデータ格納部１６３から読み込んだ要約パラメータのデフォルト値が格納される。操作者が入力する要約パラメータとしては、例えば、全文書に対する要約の比率（１～９９％）や、日付時刻、価格情報、物理量（サイズ、重量、温度等）等の数量優先のある／なし、ＵＲＬ（UniformResource Locator）重視長単文の優先のある／なし、です／ます／であるの選択をする／しない、等の値が格納される。タームベクトル格納エリア１１３４には、本実施形態により取得された、嗜好要約作成の対象文書の、タームベクトルが格納される。要約格納エリア１１３５には、本実施形態により取得された重要文が、嗜好要約作成対象文書における順番で格納される。

【００１０】キーボード１２は、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。マウス１３は、ポインティングデバイスであり、表示装置１４に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。表示装置１４は、例えばＣＲＴや液晶ディスプレイ等が使用される。この表示装置１４には、嗜好要約作成の対象となる文書の内容や、本実施形態により作成された嗜好要約等が表示されるようになっている。印刷装置１５は、表示装置１４に表示された文章や、記憶装置１６の文書データベース１６５に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットプリンタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。

【００１１】記憶装置１６は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置１６に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置１７で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。記憶装置

16は、仮名漢字変換辞書161、プログラム格納部162、データ格納部163、重要語データベース164、文書データベース165、行列データベース168、図示しないその他の格納部（例えば、この記憶装置16内に格納されているプログラムやデータ等をバックアップするための格納部）等を有している。プログラム格納部162には、本実施形態における嗜好要約作成処理プログラム等の各種プログラムの他、仮名漢字変換辞書161を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。データ格納部163には、要約パラメータのデフォルト値等の各種データが格納されている。要約パラメータのデフォルト値としては、例えば、全文書に対する要約の比率＝「25%」や、日付時刻、価格情報、物理量（サイズ、重量、温度等）等の数量重視＝「しない」や、URL（Uniform Resource Locator）重視＝「しない」、長単文の重視＝「しない」や、です／ます／であるの選択＝「しない」、等の値が格納されている。

【0012】重要語データベース164には、本実施形態において、過去の所定期間中に処理された文書をもとに取得されたキーワード（処理重要語）とこのキーワード（処理重要語）の重要度（処理重要度）が互いに対応して格納されている。文書データベース165には、仮名漢字変換プログラムにより作成された文書や、他の装置で作成されて記憶媒体駆動装置17や通信制御装置18から読み込まれた文書が格納される。この文書データベース165に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、HTML（Hyper TextMarkup Language）形式の文書、JIS形式の文書等の各種形式の文書の格納が可能である。更にこの文書データベース165には、文書を処理したユーザー及びその処理回数が各文書に対応付けて格納されている。前記処理回数は、所定期間毎に値を0にリセットされる。

【0013】行列データベース168には、過去の所定期間に行われた文書処理の処理内容により取得される行列Ga，Gb，Gcが格納されている。これらの行列Ga，Gb，GcからGP（Group Personalize ）行列が取得され、このGP行列によって、要約対象文書の重要語（句も含む）の重要度がシフト（重要度が変換）される。図2（a）～（c）は、行列Ga，Gb，Gcの一例を示す説明図である。

【0014】行列Gaは、図2（a）に示すように、過去所定期間内に処理した処理文書から抽出された処理重要語を行に、同処理文書を列にとった行列であり、各要素は処理重要語の処理重要度f（x）を表している。行列Gbは、図2（b）に示すように、前記処理文書を行にとり、ユーザーを列にとった行列であり、各要素は、ユーザーが各文書を前記所定期間内に処理した回数となっている。行列Gcは、図2（c）に示すように、行および列がともにユーザーそれぞれの重要度係数を示している。行列Ga及び行列Gbは所定期間ごとに書き換えられ、行列Gcは操作者からの入力により適宜書き換えられる。

【0015】記憶媒体駆動装置17は、CPU111が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラムには、本実施形態の文書処理装置により実行される各種処理のためのプログラム、および、そこで使用される辞書、データ等も含まれる。ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやICカード等の半導体記憶媒体、CD－ROMやMO、PD（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。本実施形態の文書処理装置において使用される記憶媒体としては、主として、CD－ROMやフロッピーディスクが使用される。記憶媒体駆動装置17は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み可能な記憶媒体に対してRAM113や記憶装置16に格納されているデータ等を書き込むことが可能である。

【0016】本実施形態の文書処理装置では、制御部11のCPU111が、記憶媒体駆動装置17にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置16の各部に格納（インストール）する。そして、本実施形態による類似度算出等の各種処理を実行する場合、記憶装置16から該当プログラムをRAM113に読み込み、実行するようになっている。但し、記憶装置16からではなく、記憶媒体駆動装置17により外部の記憶媒体から直接RAM113に読み込んで実行することも可能である。また、文書処理装置によっては、本実施形態の嗜好要約作成処理プログラム等を予めROM112に記憶しておき、これをCPU111が実行するようにしてもよい。

【0017】通信制御装置18は、他のパーソナルコンピュータやワードプロセッサ等との間でテキスト形式やHTML形式等の各種形式の文書やビットマップデータ等の各種データの送受信を行うことができるようになっている。入出力I／F19は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。文字認識装置20は、用紙等に記載された文字をテキスト形式やHTML等の各種形式で認識する装置であり、イメージスキャナや文字認識プログラム等で構成されている。

【0018】本実施形態では、キーボード12の入力操作により作成した文書（RAM113の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置17から読み込んだ文書、予め文書データベースに格納されている文書、通信制御装置18からダウンロードした文書、及び文字認識装置20で文字認識した文書、等の各種文書を対象文書として取得する（文書取得手段）ことが可能である。

【0019】次に、上述のような構成の文書処理装置による嗜好要約作成処理であって、本発明の文書処理方法の一実施形態について図3～図7を参照して説明する。

【0020】本実施形態においては、所定期間毎に、該所定期間内に行われた文書処理の処理内容基づいて新たな処理重要語及び処理重要度が取得され、行列データベース168内の行列Ga及び行列Gbが書き換えられる。

【0021】図3は、行列Ga，Gb書き換え処理の動作を表したフローチャートである。CPU111は、所定期間内に処理された文書（処理文書）を文書データベース165から順次取得してRAM113の所定作業領域に格納し（ステップ11）、各文書についてのキーワード（処理重要語（句も含む））及びその重要度（処理重要度）を取得する（ステップ12）。

【0022】図4は、各文書についての処理重要語・処理重要度取得処理の動作を表したフローチャートである。図4に示すように、CPU111は、文書データベース165から取得した文書について、形態素解析を行うことで処理文書から自立語を抽出する（ステップ121）と共に、名詞句、複合名詞句等を含めた候補語（句）を処理文書から抽出する（ステップ122）。次に抽出した候補語（句）の処理文書での出現頻度、評価関数から、各候補語（句）の処理重要度f（x）を取得する（ステップ123）。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。

【0023】さらにCPU111は、取得した処理重要度f（x）の値をもとに候補語（句）から処理重要語a，b，c，…を取得し（ステップ124）、この処理重要語a，b，c，…及びその処理重要度f（a），f（b），f（c）…を重要語データベース164に格納する。すべての処理文書について、処理重要語及びその処理重要度を取得すると、図3に示す行列Ga，Gb書き換え処理ルーチンへリターンする。

【0024】次にCPU111は、行列データベース168の行列Gaを、前記処理重要語a，b，c，…を行に、前記所定期間の処理文書を列に、また処理重要度f（x）を各要素にとったものに書き換える（ステップ13）。そして、CPU111は、文書データベース165から、各文書の処理回数を取得し（ステップ14）、行列Gbを、所定期間内の処理文書を行に、文書データベース165から取得した処理回数を各要素としたものに書き換えて、行列Ga，Gb書き換え処理を終了する。

【0025】図5は、嗜好要約作成処理のメイン動作を表すフローチャートである。要約作成処理に際しては、CPU111は、要約を作成する対象となっている文書（要約対象文書）を取得し、RAM113の対象文書格納エリア1131に格納する（ステップ21）。要約対象文書は、ユーザの指示に従ってRAM113、記憶装置16の文書データベース165、記憶媒体駆動装置17、または通信制御装置18から取得する。続いてCPU111は、ユーザによってキーボード12等から要約パラメータが入力された場合には入力値を取得し、ユーザによる入力がない場合にはデータ格納部163に格納された要約パラメータのデフォルト値を取得し、要約パラメータ格納エリア1132に格納する（ステップ22）。

【0026】次にCPU111は、対象文書格納エリア1131に格納した要約対象文書に対するタームベクトルVを求める（ステップ23）。図6は、タームベクトル取得処理の動作を表したフローチャートである。CPU111は、まず形態素解析を行うことで要約対象文書に含まれる自立語を抽出する（ステップ231）と共に、名詞句、複合名詞句等を含めた候補語（句）を要約対象文書から抽出しRAM113の所定作業領域に格納する（ステップ232）。次に、RAM16の要約パラメータ格納エリア1132に格納した要約パラメータや、抽出した候補語（句）の要約対象文書中での出現頻度、評価関数等から、客観的な重要度g（y）を決定する（ステップ233）。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。

【0027】そして、この客観的な重要度g（y）により重要語p，q，r，…を取得し（ステップ234）、重要語p，q，r，…の客観的な重要度g（p），g（q），g（r），…を要素とするタームベクトルVを取得し（ステップ235）、図5に示す嗜好要約作成処理のルーチンへリターンする。

【0028】続いて、CPU111は、行列Gaを行列データベース168から取得し、タームベクトルVと行列Gaとの次元合わせを行う（ステップ24）。即ち、タームベクトルVの次元数と、行列Gaの行数とを、要約対象文書の重要語と行列Gaの行があらわす処理重要語の和集合の数とし、タームベクトルVのみに含まれる重要語に対する行列Gaの要素値、および、行列の行のみに含まれる重要語に対するタームベクトルVの要素値は、”0”と定義する。

【００２９】例えば、要約対象文書の重要語が「重要、重要語、重要度、…」、行列Ｇａの行があらわす処理重要語が「重要、…、政治、…」であり、要約対象文書のタームベクトルＶ＝（　１，１８，１９，…）、行列Ｇａの、ある１列が（１８，…，２１，…）である場合、次元を合わせると、要約対象文書のタームベクトルＶ＝（　１，１８，１９，…，　０，…）、行列Ｇａの１列は（１８，　０，　０，…，２１，…）となる。次元合わせ後の行列Ｇａ及びタームベクトルＶは、それぞれ、ＲＡＭ１１３の行列格納エリア１１３５、タームベクトル格納エリア１１３４に格納する。

【００３０】続いて、ＣＰＵ１１１は、行列Ｇｂ，Ｇｃを行列データベース１６８から取得し、次元を合わせを行った行列Ｇａと行列Ｇｂ，ＧｃとからＧＰ行列を取得する（ステップ２５）。ＧＰ行列は、次の式に従って求める。ＧＰ＝Ｇａ・Ｇｂ・Ｇｃ従って、本実施形態におけるＧＰ行列は、Ｇａ行列の次元合わせを行った行をそのまま行にとり、ユーザーの各メンバーを列にとってなっており、ＧＰ行列の各要素は、メンバー毎の過去の文書処理における処理重要語の処理重要度ｆ（ｘ）に各メンバーの重要度を加味して表した数値となっている。

【００３１】ＧＰ行列が取得されると、続いてＣＰＵ１１１は、このＧＰ行列をもとにＧＰベクトルを算出する（ステップ２６）。

【００３２】図7は、ＧＰ行列からＧＰベクトルを算出する行程を概念的に説明する説明図である。ＣＰＵ１１１は、まず、ＧＰ行列の各要素ｇｉｊ（ｉ＝1～メンバー数ｍ、ｊ＝1～要約対象文書の重要語と処理重要語の和集合の数ｋ）の各行毎の要素の平均値を算出して列ベクトル（総ＧＰベクトル）を得る（図7（１）→（２））。この総ＧＰベクトルは、各要素ｇｉが重要語毎のユーザーグループ全体における過去の文書処理での出現頻度（但し各重要語の予め決められた重要語の重み等や、メンバーの重要度が加味されている）を反映した数値となっている。ＣＰＵ１１１は、更に、この総ＧＰベクトルの各要素ｇｉを文書の処理回数の総数で割って、１列のＧＰベクトルを得る（図7（２）→（３））。この様に、総ＧＰベクトルを文書の処理回数の総数で割るのは、行列Ｇｂに文書の処理回数が要素として含まれており、処理回数が増えるに従ってＧＰベクトルが大きくなっていくのを回避するためである。

【００３３】そして、ＣＰＵ１１１は、ＧＰベクトルの各要素とこの各要素に対応するタームベクトルＶの要素とを掛け合わせて、嗜好タームベクトルＶ’を得る（ステップ２７）。この嗜好タームベクトルＶ’の各要素は、客観的な重要度ｇ（ｙ）にユーザーのタームについての嗜好を重み付けした嗜好重要度ｇ’（ｙ）となっている。

【００３４】続いて、ＣＰＵ１１１は、重要語の嗜好重要度ｇ’（ｙ）により、要約対象文書に含まれる嗜好部分重要度（嗜好文重要度Ｆ（Ｚ））を取得する（ステップ２８）。そして、決定した各部分（各文）の嗜好部分重要度（嗜好文重要度Ｆ（Ｚ））の高い部分（文）の上位から要約パラメータの要約比率（例えば、対象要約文書中の全文数の内の上位２５％）以内に入る部分（文）を嗜好重要部分（嗜好重要文）としてリストアップし（ステップ２９）、リストアップした文を要約対象文書の中での出現順に並べることで当該要約対象文書の嗜好要約とし、これをＲＡＭ１１３の要約格納エリア１１３７に格納して（ステップ３０）、本実施形態による嗜好要約作成処理を終了する。

【００３５】この様に、本実施形態では、過去の文書処理における出現頻度等をもとにユーザーの重要語に対する嗜好を把握し、要約対象文書から取得した重要語の客観的な重要度ｇ（ｙ）を前記ユーザーの嗜好を反映して重み付けをした嗜好重要度ｇ’（ｙ）に変換し、この嗜好重要度ｇ’（ｙ）をもとに重要文を取得して要約を作成する。従って、本実施形態によると、ユーザーの嗜好の反映された要約が作成される。本実施形態によると、重要語の客観的な重要度を要素としたタームベクトルＶを獲得し、ユーザーの嗜好を反映させたＧＰ行列を用いて変換させることによって、嗜好重要度を要素とする嗜好タームベクトルＶ’を獲得しているので、計算処理が簡単であり、ベクトル空間法を採用したコア・エンジンを備えた一般の文書処理装置に容易に適用することが可能である。

【００３６】本実施形態よると、タームベクトルＶを嗜好タームベクトルＶ’にシフトさせるＧＰ行列を、表現すべき特徴毎の単純な観点で構成した行列Ｇａ，Ｇｂ，Ｇｃの掛け合わせて求めているので、様々な特徴を考慮に入れたＧＰ行列を容易に構成してタームベクトルＶをシフトさせることが可能である。本実施形態よると、タームベクトルＶを嗜好タームベクトルＶ’にシフトさせるためのＧＰ行列は、各列がユーザーの興味を反映しているので、複数のユーザーからなるグループを数グループに分割した該グループのＧＰ行列や個々のユーザーのＧＰ行列（ベクトル）を容易に得ることができる。本実施形態よると、ＧＰ行列がユーザーの過去に処理した文書をもとに所定期間毎に書き換えられている行列Ｇａ，Ｇｂ，Ｇｃをもとに取得されているので、タームベクトルＶがユーザーの嗜好の経時的変化に対応した嗜好タームベクトルＶ’にシフトされ、ユーザーの嗜好の変遷に追随した嗜好要約が作成される。

【００３７】尚、本発明は、上述の実施形態に限定されるものではなく、本発明の趣旨を逸脱しない限りにおいて適宜変更が可能である。上述の実施形態においては文書処理装置としてコンピュータを用いているが、コンピュータに限定されるものではなく、ワードプロセッサ等であってもよい。

【００３８】要約対象文書から取得した重要語候補すべ

てについて嗜好重要度を獲得し、この嗜好重要度に基づいて重要語候補から重要語を取得することもできる。客観的な重要度に基づいて嗜好重要度を取得する場合に、客観的な重要度をベクトル化せずに、客観的な重要度それぞれに適当な処理を施すことにより嗜好重要度を得ることもできる。また、客観的な重要度をベクトル化する場合であっても、タームベクトルを嗜好タームベクトルに変換する手法はＧＰ行列を用いていなくてもよい。候補語抽出手段及び重要語獲得手段として、要約作成対象文書から処理重要語を抽出する処理重要語抽出手段を用いることもできる。

【００３９】上述の実施形態においてはＧＰ行列は、ユーザー一人ずつの過去の文書処理回数（行列Ｇａ）と各文書における重要語の出現頻度（行列Ｇｂ）、および各ユーザーの重要度（行列Ｇｃ）とから取得されているが、ユーザー毎の過去の文書処理回数（行列Ｇａ）と各文書における重要語の出現頻度（行列Ｇｂ）のみにより取得されてもよい。また、例えば、各文書の処理時間や、他の文書作成に引用された件数等も加味して取得されてもよい。更に、ＧＰ行列を上述の実施形態と同様に行列Ｇａ～行列Ｇｃ等の行列の掛け合わせから取得する場合において、行列Ga ～行列Ｇｃ等の各行列の要素はそれぞれ重要語の文書中の出現頻度や、ユーザーが各文書を処理した回数を反映した数値となっていればよく、直接出現頻度や処理回数そのものを表していなくてもよい。

【００４０】上述の実施形態においては行列Ｇａ～Ｇｃは過去の文書処理内容から取得されているが、ユーザーから取得して行列データベース１６８に格納しておいてもよい。上述の実施形態においては行列Ｇａ～Ｇｃは所定期間毎に書き換えられているが、文書処理毎にまたは操作者等の判断により適宜書き換えるようにしてもよい。ＧＰベクトルを表示装置に表示するＧＰベクトル表示手段を備え、ユーザーのグループ全体やユーザーの嗜好を視覚的に把握できるようにしてもよい。この場合、ＧＰベクトルを行列データベースまたは専用のＧＰベクトルデータベースに経時順に格納しておき、経時変化も把握できるようにしてもよい。上述の実施形態においては、重要語句の嗜好重要度によって文単位で重要度が比較され、嗜好重要部分として嗜好重要文が選択されるが、段落単位やタイトルの重要度を比較して、嗜好重要部分として嗜好重要段落や嗜好重要タイトルを選択させるようにしてもよい。

【００４１】

【発明の効果】以上説明したように、本発明によれば、要約対象文書中の重要語について、ユーザーの嗜好を踏まえた嗜好重要度を取得し、この嗜好重要度にもとづいて重要部分を選択し、この重要部分から要約を作成するので、作成された要約にユーザーの興味や注目度、目的等の嗜好が反映される。

【図面の簡単な説明】

【図１】本発明の文書処理装置の一実施形態であり、本発明の文書処理プログラムを記憶した記憶媒体の一実施形態の該プログラムが読み取られたコンピュータの構成を表したブロック図である。

【図２】図１の実施形態における行列Ｇａ，Ｇｂ，Ｇｃの一例を示す説明図である。

【図３】図１の実施形態における行列Ｇａ，Ｇｂ書き換え処理の動作を表したフローチャートである。

【図４】図１の実施形態における、各文書についての処理重要語・処理重要度取得処理の動作を表したフローチャートである。

【図５】図１の実施形態における嗜好要約作成処理のメイン動作を表すフローチャートである。

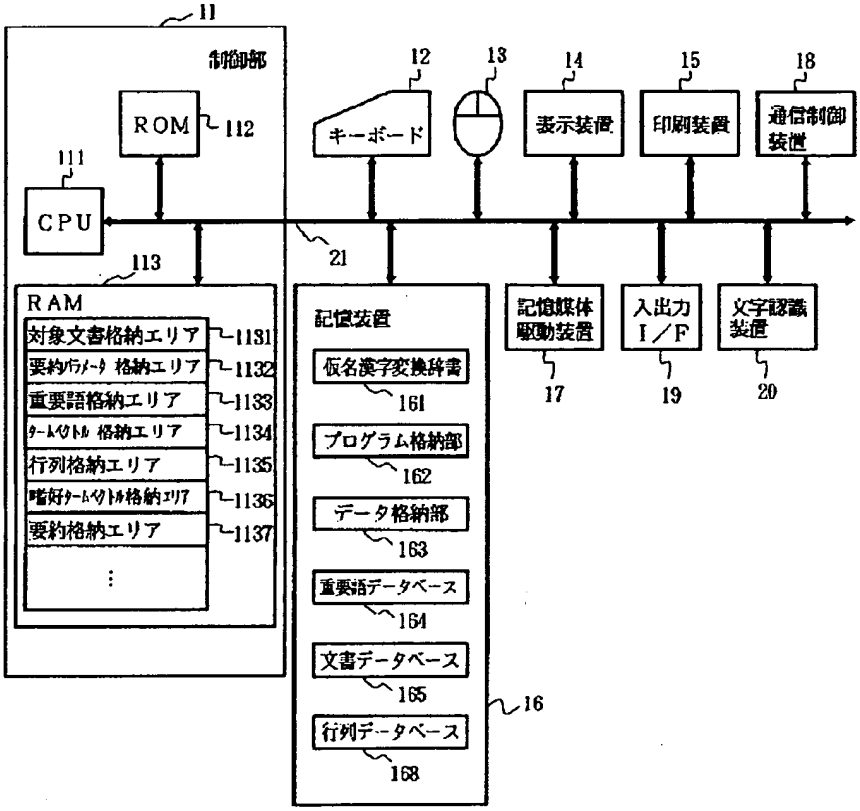
【図６】図１の実施形態におけるタームベクトル取得処理の動作を表したフローチャートである。

【図７】図１の実施形態においてＧＰ行列からＧＰベクトルを取得する行程を概念的に説明する説明図である。

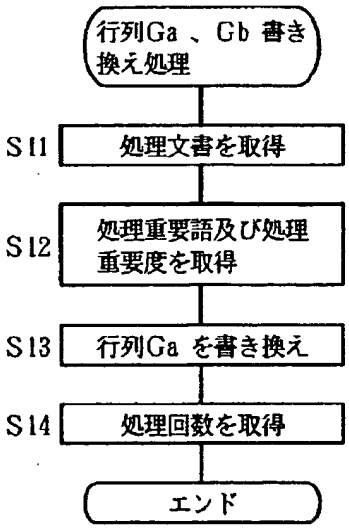
【符号の説明】

１１　制御部
１１２　ＲＯＭ
１１３　ＲＡＭ
１１３１　対象文書格納エリア
１１３２　要約パラメータ格納エリア
１１３３　重要語格納エリア
１１３４　タームベクトル格納エリア
１１３５　行列格納エリア
１１３６　嗜好タームベクトル格納エリア
１１３７　要約格納エリア
１２　キーボード
１３　マウス
１４　表示装置
１５　印刷装置
１６　記憶装置
１６１　仮名漢字変換辞書
１６２　プログラム格納部
１６３　データ格納部
１６４　重要語データベース
１６５　文書データベース
１６８　行列データベース

【図1】


11
制御部
ROM
112
111
CPU
113
RAM
対象文書格納エリア
1131
要約パラメータ 格納エリア
1132
重要語格納エリア
1133
タームベクトル 格納エリア
1134
行列格納エリア
1135
嗜好タームベクトル格納エリア
1136
要約格納エリア
1137
12
キーボード
13
14
表示装置
15
印刷装置
18
通信制御装置
21
記憶装置
仮名漢字変換辞書
161
プログラム格納部
162
データ格納部
163
重要語データベース
164
文書データベース
165
行列データベース
168
16
記憶媒体駆動装置
17
入出力I/F
19
文字認識装置
20

【図3】


行列Ga 、Gb 書き換え処理
S11
処理文書を取得
S12
処理重要語及び処理重要度を取得
S13
行列Ga を書き換え
S14
処理回数を取得
エンド

【図2】

（a）　行列　Ga（処理重要語、処理文書）

	文書A	文書B	文書C	
⋮	⋮	⋮	⋮	
重　要	2	1	18	…
重要語	20	18	0	…
重要度	21	19	0	…
⋮	⋮	⋮	⋮	…
秋　治	0	0	21	…
⋮	⋮	⋮	⋮	

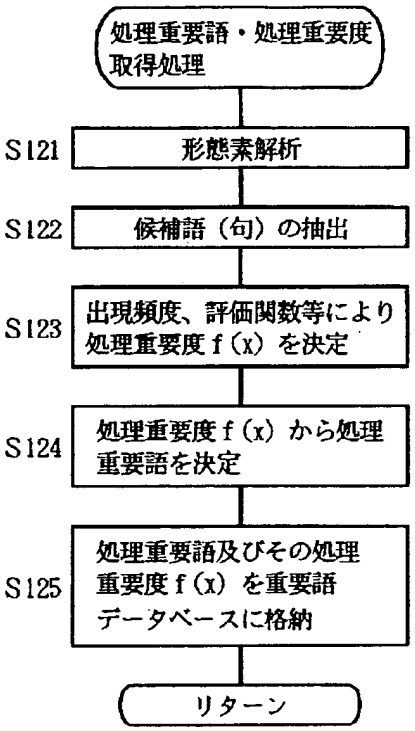
（b）　行列　Gb（処理文書、ユーザー）

	星太郎	花園美子	黒帯三四郎	鹿見五郎
文書A	1	0	1	0
文書B	1	1	2	0
文書C	1	1	1	1
⋮	⋮	⋮	⋮	⋮

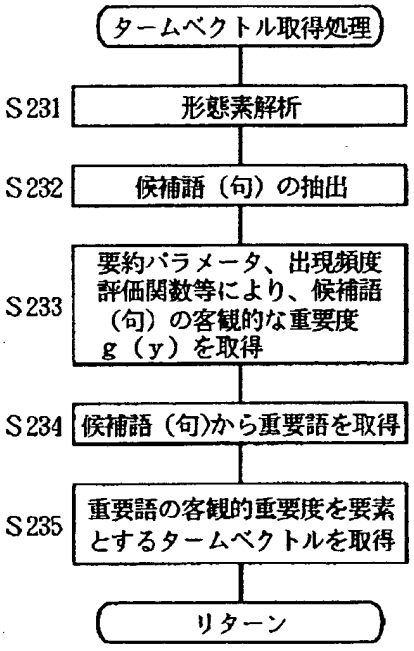
（c）　行列　Gc（ユーザーの重要度）

	星太郎	花園美子	黒帯三四郎	鹿見五郎
星太郎	1.5	0	0	0
花園美子	0	0.8	0	0
黒帯三四郎	0	0	1.3	0
鹿見五郎	0	0	0	1.1

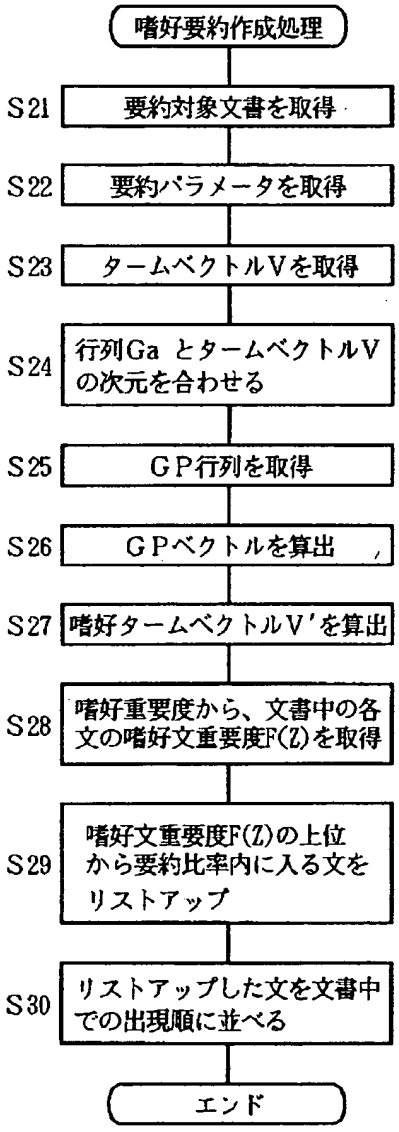
【図4】



【図6】



【図5】


嗜好要約作成処理
S21
要約対象文書を取得
S22
要約パラメータを取得
S23
タームベクトルVを取得
S24
行列Ga とタームベクトルV の次元を合わせる
S25
GP行列を取得
S26
GPベクトルを算出
S27
嗜好タームベクトルV'を算出
S28
嗜好重要度から、文書中の各文の嗜好文重要度F(Z)を取得
S29
嗜好文重要度F(Z)の上位から要約比率内に入る文をリストアップ
S30
リストアップした文を文書中での出現順に並べる
エンド

【図7】

メンバー

$$\text{重要語}\begin{pmatrix} g_{11} & g_{12} & g_{13} & \cdots & g_{1m} \\ g_{21} & g_{22} & g_{23} & \cdots & g_{2m} \\ g_{31} & g_{32} & g_{33} & \cdots & g_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ g_{k1} & g_{k2} & g_{k3} & \cdots & g_{km} \end{pmatrix} \qquad (1)$$

GP行列

↓ 平均化

$$\begin{pmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_k \end{pmatrix} \qquad (2)$$

$$\left(g_i = \frac{g_{i1} + g_{i2} + \cdots g_{im}}{m} \; ; \; i = 1 \sim k \right)$$

総GPベクトル

↓ 規格化

$$\begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ \vdots \\ h_k \end{pmatrix} \qquad (3)$$

$$\left(h_i = \frac{g_i}{\text{各文書の処理回数の合計}} \right)$$

GPベクトル